Tìm hiểu về Spark

1. Spark properties
2. Spark RDD
3. Spark DataFrame

1) Spark properties

Spark properites kiểm soát hầu hết các cài đặt của chương trình và có thể hiệu chỉnh riêng biệt ở mỗi chương trình. Các thiết lập được cài đặt trong kiểu dữ liệu SparkConf và được lưu đến trong SparkContext.

Trong PySpark, lớp SparkContext gồm có các giá trị:

* master: URL của cluster mà chương trình kết nối.
* appName: tên chương trình.
* sparkHome: vị trí spark được cài đặt.
* pyFiles: file .zip hay .py cần được thêm vào môi trường.
* enviroment: nút biến môi trường
* …

Trong các giá trị trên, master và appName là các giá trị thường được gọi làm parameter khi khởi tạo:

from pyspark import SparkContext

sc = SparkContext("local", "BigData\_Homework2")

2) Spark RDD

Spark RDD(Resilient Distributed Datasets), bộ dự liệu linh hoạt, là một cấu trúc dữ liệu của Spark, bao gồm các tập đối tượng cố định và được phân loại, một khi khởi tạo dữ liệu không thể thay đổi. Mỗi tệp dữ liệu thuộc RDD được phân loại dựa trên các phép logic. Có 2 cách để tạo ra RDD: parallelizing, một tập dữ liệu có sẵn trong chương trình và referencing a dataset, tham chiếu một tập dữ liệu trong bộ nhớ ngoài.

Trong pyspark, để khởi tạo một kiểu dữ liệu RDD:

from pyspark import SparkContext

sc = SparkContext("local", "BigData\_Homework2")

rdd = sc.parallelize(['TDTU','IT','BigData'])

3) Spark DataFrame

DataFrames là các tập hợp dữ liệu phân tán bất biến, trong đó dữ liệu được sắp xếp theo cách quan hệ - nghĩa là, các cột được đặt tên tương tự như các bảng trong cơ sở dữ liệu quan hệ. Bản chất của bộ dữ liệu là áp dụng cấu trúc trên bộ sưu tập dữ liệu phân tán để cho phép xử lý hiệu quả và dễ dàng hơn.